|  |  |
| --- | --- |
| **Título:**  Uso de big data em saúde no Brasil: perspectivas para um futuro próximo | |
| **Local de Publicação:**  Universidade de São Paulo, Faculdade de Saúde Pública, São Paulo-SP, Brasil  doi: 10.5123/S1679-49742015000200015 | |
| **Responsável pelo fichamento:**  Maurício Collaça Ramos | |
| **Autor principal (nome e link para o Lattes ou outra fonte):** | Alexandre Dias Porto Chiavegatto Filho  <http://lattes.cnpq.br/5517850224634709> |
| **Tema\Assunto:** | O uso de big data na área da Saúde no Brasil, em especial a epidemiologia, contemplando linguagens de programação, técnicas estatísticas e a questão da privacidade de dados confidenciais. |
| **Abordagem metodológica utilizada:** | Pesquisa exploratória com abordagem qualitativa na análise de documentos. |
| **Principais conceitos, terminologia e enfoques**  **que sustentam a argumentação:** | Big data – quando a complexidade e a quantidade de dados são grandes o bastante para transformar a forma como os dados são analisados tradicionalmente.  Medicina de precisão – é aquela em que medicamentos são prescritos para pacientes específicos e não para a população de pacientes só porque sejam estatisticamente significativos.  PEP – O Prontuário Eletrônico do Paciente pode ser usado de forma integrada por todos os estabelecimentos de saúde, poupando tempo de preenchimento, reduzindo o viés de memória/esquecimentos, aumentando a completude e o uso potencial em pesquisa de maior abrangência. No Brasil que possui uma atuação forte da saúde privada será necessária liderança do SUS face a complexidade.  Internet das Coisas - é uma rede de objetos físicos, veículos, prédios e outros que possuem tecnologia embarcada, sensores e conexão e é com rede capaz de coletar e transmitir dados. Especificamente quanto a saúde, sensores poderão identificar quedas (idosos) ou eminência de infartos e acidentes vasculares antes do próprio indivíduo. Estes sensores poderão gerar infinita quantidade de dados para pesquisas, permitindo identificar os fatores imediatos ou distantes que levam pessoas a doenças ou óbitos.  Linguagens de programação para análise de dados – Fundamentais para o compartilhamento de resultados e desenvolvimentos de novas análises, é importante que haja uma padronização entre os pesquisadores. Atualmente R e Python são as mais difundidas, nota-se um uso crescente da Julia e outras linguagens já muito utilizadas são Stata, SAS, SPSS, JMP, MATLAB.  Machine Learning - O aprendizado de máquina é um subcampo da ciência da computação que evoluiu do estudo de reconhecimento de padrões e da teoria do aprendizado computacional em inteligência artificial. É o campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados, explorando o estudo e construção de algoritmos que podem aprender de seus erros e fazer previsões sobre dados.  Associações espúrias – problema que se agrava em Big Data, são quando variáveis possuem uma associação aleatória entre elas e no caso de centenas ou milhares de testes de hipóteses simultâneos a possibilidade de serem consideradas significantes é enorme, e que são historicamente corrigidos com métodos chamados correção de Bonferroni ou taxa de falsas descobertas (FDR), estes últimos controlados por abordagens Benjamini-Hochberg (BH) e Storey.  Redução da dimensão – métodos estatísticos que visam reduzir a dimensão dos dados, permitir melhor visualização dos dados, diminuir a quantidade de memória RAM necessária para rodar os modelos, diminuir a quantidade de associações espúrias através da transformação de variáveis correlacionadas em não-correlacionadas e melhorar modelos preditivos. No Big Data e na epidemiologia os métodos com mais atenção recente são a análise dos componentes principais (ACP) e o *propensity score*. |
| **Fatos históricos, dados e outros autores que apoiam a argumentação:** | Nos últimos anos encontrou-se que a próxima grande fronteira da epidemiologia será o big data. O aumento das pesquisas multicêntricas e a cobrança da prestação de contas dos gastos públicos aumenta a quantidade de dados e demanda por análises de dados complexos.  Sistema de Informações sobre Nascidos Vivos (SINASC) desde 1990,  o Sistema de Informações sobre Mortalidade (SIM) desde 1975 são exemplos de oportunidades de dados para big data na área de Saúde.  Na metade do século passado, encontrar os parâmetros de uma regressão linear com 500 observações era uma tarefa que levava alguns dias. Hoje, são necessários também alguns dias para rodar modelos bayesianos com centenas de milhares de observações. |
| **Ideias centrais, proposições ou teses defendidas:** | Crescimento do uso de big data nos últimos anos, incluindo todas as áreas científicas.  Áreas de aplicação do big data em saúde são a medicina de precisão, os prontuários eletrônicos (PEP) como evolução do cartão SUS e a Internet das Coisas com sua rede de sensores em objetos e indivíduos que permitirá o acionamento automático de sistemas de saúde em caso de adversidades bem como facilitar a coleta de dados em larga escala para monitoramento e diagnóstico de doenças, agravos e óbitos.  Para a análise de dados, data-mining e big data, diversas linguagens vem sendo utilizadas mas nos últimos anos um consenso vem se formando em torno do R e do Python.  Espera-se maior utilização de aprendizado de máquina para modelos de predição de classificação e regressão com uso de métodos de controle de associações espúrias tais como correção Bonferroni ou controle da taxa de falsos negativos (FDR) pelos métodos Benjamini-Hochberg e Storey.  Crescente importância da questão da privacidade de dados dos pacientes, cujo vazamento pode ocorrer por descuido do cientista ou invasão proposital.  O uso de big data na área da saúde auferirá ganhos importantes recurso de tempo e dinheiro, em vidas e deve ser defendido e comunicado à população pela comunidade científica envolvida. |
| **Ideias secundárias, proposições ou teses que**  **complementam o pensamento do autor:** | N/A |
| **Outros artigos e publicações do mesmo autor:** | Diaz-Quijano, F. ; Pelissari, D. M. ; Chiavegatto Filho, A.D.P.. Zika-associated microcephaly epidemic and birth rate reduction in Brazilian cities. AMERICAN JOURNAL OF PUBLIC HEALTH, v. 01, p. e1-e3, 2018.  Santana, G. L. ; Coelho, B. M. ; Wang, Y.P. ; Chiavegatto Filho, A.D.P. ; Viana, M.C. ; Andrade, L.H.S.G.. The epidemiology of personality disorders in the Sao Paulo Megacity general population. PLoS One, v. 13, p. e0195581, 2018.  Diaz-Quijano, F. ; Chiavegatto Filho, A.D.P.. Reduction of the birth rate in São Paulo: a probable effect of the panic caused by the Zika-associated microcephaly epidemic. ANNALS OF EPIDEMIOLOGY, v. 28, p. online first, 2017.  Wang, Y.P. ; Chiavegatto Filho, A.D.P. ; Campanha, A. ; Malik, A.M. ; Mogadouro, M. ; Cambraia, M. ; Viana, M. ; Andrade, L.H.S.G. . Patterns and predictors of health service use among people with mental disorders in São Paulo metropolitan area, Brazil. Epidemiology and Psychiatric Sciences, v. 1, p. 1-13, 2016.  Nunes B.P. ; Flores, T. R. ; Garcia, L. P. ; Chiavegatto Filho, A.D.P. ; Facchini, L. A. . Tendência temporal da falta de acesso aos serviços de saúde no Brasil, 1998-2013. Epidemiologia e Serviços de Saúde, v. 25, p. 777-787, 2016.  Chiavegatto Filho, A.D.P.; Kawachi, I. . Income inequality is associated with adolescent fertility in Brazil: a longitudinal multilevel analysis of 5,565 municipalities. BMC Public Health (Online), v. 15, p. 103-107, 2015.  Chiavegatto Filho, A.D.P.; Wang, Y.P. ; Malik, A.M. ; Takaoka, J. ; Viana, M.C. ; Andrade, L.H.S.G. . Determinants of the use of health care services: multilevel analysis in the Metropolitan Region of Sao Paulo. Revista de Saúde Pública (Impresso), v. 49, p. 1-12, 2015.  Chiavegatto Filho, A.D.P.; Kawachi, I. ; Wang, Y.P. ; Viana, M.C. ; Andrade, L.H.S.G. . Does income inequality get under the skin? A multilevel analysis of depression, anxiety and mental disorders in São Paulo, Brazil. Journal of Epidemiology and Community Health (1979), v. 67, p. 966-972, 2013.  Chiavegatto Filho, A.D.P.; Gotlieb, S. L. D. ; de Almeida, S. L. ; Kawachi, I. . Como incluir características dos distritos do município de São Paulo em estudos epidemiológicos? Análise da desigualdade de renda pelo uso do propensity score matching. Saúde e Sociedade (USP. Impresso), v. 22, p. 1145-1153, 2013.  Chiavegatto Filho, A.D.P.; Kawachi, I. ; Lebrão, M. L. . Income inequality and elderly self-rated health in São Paulo, Brazil. Annals of Epidemiology, v. 22(12), p. 863-867, 2012.  Chiavegatto Filho, A.D.P.; Kawachi, I. ; Gotlieb, S.L.D. . Propensity score matching approach to test the association of income inequality and mortality in Sao Paulo, Brazil. Journal of Epidemiology and Community Health (1979), v. 66, p. 14-17, 2012.  de Mello Jorge, M.H.P. ; Laurenti, R. ; LIMA-COSTA, M. F. ; Gotlieb, S. L. D. ; Chiavegatto Filho, A.D.P. . A mortalidade de idosos no Brasil: a questão das causas mal definidas. Epidemiologia e Serviços de Saúde, v. 17, p. 271-281, 2008.  Chiavegatto Filho, A.D.P.; Laurenti, R. ; Gotlieb, S. L. D. ; de Mello Jorge, M.H.P. . Desnutrição como causa básica ou associada de morte: análise da qualidade de informação em mulheres em idade fértil. Revista Brasileira de Epidemiologia (Impresso), v. 10, p. 30-38, 2007.  Chiavegatto Filho, A.D.P.; Laurenti, R. ; Gotlieb, S.L.D. ; de Mello Jorge, M.H.P. . Mortalidade por doença hipertensiva em mulheres de 20 a 49 anos no Município de São Paulo, SP, Brasil. Revista Brasileira de Epidemiologia (Impresso), v. 7, p. 252-258, 2004. |
| **Questões para discussão:** | Um desafio para a implementação do Big Data em Saúde é a questão da privacidade seja por descuido do cientista ou por invasões propositais. Como conscientizar os cientistas sobre a importância do desenvolvimento de protocolos de segurança mais rígidos, o uso de criptografia, anonimização do paciente e uso de terminais de acesso restrito.  Como o cientista também pode informar a população sobre os imensos ganhos de tempo, dinheiro e vidas que a análise de big data traz com o engajamento da Internet das Coisas. |